DOCUMENT RESUME

ED 332 636 HE 024 615

AUTHOR	Klitzner, Michael; Stewart, Kathryn
TITLE	Evaluating Faculty Development and Clinical Training Programs in Substance Abuse: A Guide Book.
INSTITUTION	Pacific Inst. for Research and Evaluation, Walnut Creek, CA.
SPONS AGENCY	National Inst. on Alcohol Abuse and Alcoholism (DHHS), Rockville, Md.; National Inst. on Drug Abuse (DHHS/PHS), Rockville, Md.
REPORT NO	RP0778
PUB DATE	Jun 90
NOTE	33p.
PUB TYPE	Guides - Non-Classroom Use (055) -- Reports - Descriptive (141)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Alcohol Abuse; Allied Health Occupations Education; *Clinical Teaching (Health Professions); Drug Abuse; *Evaluation Methods; *Faculty Development; Graduate Medical Education; Higher Education; Medical Education; Medical School Faculty; Mental Health; *Program Evaluation; Qualitative Research; *Research Methodology; Sampling; Statistical Bias; *Substance Abuse

ABSTRACT

Intended to provide an overview of program evaluation as it applies to the evaluation of faculty development and clinical training programs in substance abuse for health and mental health professional schools, this guide enables program developers and other faculty to work as partners with evaluators in the development of evaluation designs that meet the specialized needs of faculty development and clinical training programs. Section I discusses conceptual issues in program evaluation, including the uses of evaluation (management and monitoring, program description, program improvement, accountability, and creating new knowledge); the major options (formative/summative, process/outcome/impact, quantitative/qualitative); and the benefits and risks of conducting evaluation studies. Section II, an introduction to research methods, includes the following discussions: sampling with known sampling errors (simple random, systematic, multistage random, stratified, cluster, stratified cluster, and sequential sampling); sampling without known sampling errors (convenience, quota, modal, purposive, and snowball sampling); sample size and generalizability and sample size and statistical power; the validity of evaluations and potential sources of bias, including issues related to internal validity (history, maturation, testing, instrumentation, statistical regression, selection, mortality, interactions with selection, and ambiguity about the direction of causal influence); comparison and control groups; measurement of outcomes; and qualitative evaluation methods and analysis. (5 references) (JB)

HE

ED332636

BEST COPY AVAILABLE

EVALUATING FACULTY DEVELOPMENT AND CLINICAL TRAINING PROGRAMS IN SUBSTANCE ABUSE: A GUIDE BOOK

Michael Klitzner, Ph.D.
Kathryn Stewart, M.S.

Pacific Institute for Research and Evaluation
Bethesda, Maryland

June, 1990

Development of this *Guide* was supported by the National Institute on Alcohol Abuse and Alcoholism and the National Institute on Drug Abuse

HE 024 615

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

RP0778

ACKNOWLEDGMENTS

This *Guide Book* is based on materials originally assembled for *A Guide to First DUI Offender Program Evaluation* prepared by the authors for the California Office of Traffic Safety. The authors wish to acknowledge the contribution of Paul Moberg, Ph.D., with whom the authors originally developed the approach to program evaluation reflected in this *Guide Book*.

TABLE OF CONTENTS

TABLE OF CONTENTS (CON'T)

TABLE OF CONTENTS (CON'T)

TABLE OF CONTENTS (CON'T)

ABOUT THIS *GUIDE BOOK*

There are many text books on program evaluation and this *Guide Book* is not intended to be one of them. Instead it is intended to provide an overview of program evaluation as it applies to the evaluation of faculty development and clinical training programs in substance abuse for health and mental health professional schools. The Guide Book references other resources that can be consulted for more detailed and technical information.

The *Guide Book* is divided into two major sections. Section I discusses conceptual issues in program evaluation, including the uses of evaluation, the major options, and the benefits and risks of conducting evaluation studies. Section II is intended as an introduction to research methods. It includes discussions of sampling, statistical power, the validity of evaluations and potential sources of bias, comparison and control groups, and measurement of outcomes. A section on data analysis is not included because this aspect of evaluation has become so complex in recent years that no simple treatment is possible. However, all universities have departments of statistics that may be called upon for assistance in data analysis.

It is important to note that it is not expected that an individual unfamiliar with program evaluation will be able to plan an evaluation based on the material presented in this *Guide Book*. Rather, the purpose of the *Guide* is to enable program developers and other faculty to work as partners with evaluators in the development of evaluation designs that meet the specialized needs of faculty development and clinical training programs.

SECTION I:
ISSUES IN PROGRAM EVALUATION

USES OF PROGRAM EVALUATION

In many ways, "*program evaluation*" is a misleading phrase. "Evaluation" implies judgement of what is good and what is not good. Although determination of whether or not a program is good" is the goal of some types of program evaluation, it is by no means the only -- or even the major -- goal of program evaluation in general. Unfortunately, the "valuative" dimension of program evaluation is what most often comes to mind for many people. Thus, individuals unfamiliar with the wide range of activities subsumed under program evaluation may approach this topic with some trepidation -- especially if it is their programs that are to be evaluated.

A much more general goal of program evaluation -- and one which underlies almost all evaluation activity -- is the production of systematic information to be used in making decisions about programs. The complexity and scope of these decisions can vary widely, as can the complexity and scope of the evaluation, but the principle is always the same:

Evaluations are implemented to improve decision-making.

All the diverse, specific purposes of a given evaluation derive from this principle. With this general principal in mind, we now examine some of the specific purposes of program evaluation. In each case, the discussion will be tied to the kinds of decisions that underlie each specific purpose.

Before discussing the specific purposes of program evaluation, it is important to define clearly what is meant by a "program." In the area of faculty development and clinical training in substance abuse, a program may be as broad as an entire institution's attempt to improve its instructional program or as narrow as the techniques used by an individual practitioner to confront substance abusing patients or counsel adolescents. A program may be the introduction of new content into an existing course, or the implementation of an innovative teaching strategy. All of these "programs" may be evaluated on many different levels and for a variety of purposes. Of course, some kinds of evaluations will be more appropriate for some programs than for others. In general, however, all of the types of evaluation described below are applicable, to some degree, to any aspect of faculty development or clinical training programs.

Evaluation for Management and Monitoring

One common use of program evaluation is the monitoring and management of programs. Especially in large programs such as the implementation of a new substance abuse curriculum or the opening of a new substance abuse unit in a teaching hospital, so much is going on that it is often difficult to keep track of the program on a day-to-day basis.

Management information or monitoring systems are designed to meet ongoing information needs (i.e., those information needs that can be expected to exist as long as the program is in operation). In general, the questions addressed by such systems relate to events that are recurrent and routine. Examples of questions related to an institution's instructional program might include:

How many hours of didactic and clinical instruction on substance abuse are offered each quarter or semester?

How many students are exposed each quarter or semester to a given course or clinical experience?

What are the trends in student exam scores in the area of substance abuse?

Similarly, questions concerning the provision of clinical services might include:

How many patients are diagnosed each month with substance-related problems? On what services are they diagnosed? Where are they referred?

How many beds are filled at any given time on the substance abuse unit? From where were patients referred? What are the intake diagnoses?

Regular review of the data generated by a management information or monitoring system provides an up-to-date picture of the program as a whole or of specific program elements (e.g., a specific course or practicuum) which can be used both to make short-term corrections and to plan future program directions.

Evaluation for Program Description

Often, there is a need to capture the overall operation of a program or program element at a given point in time -- that is, to develop a comprehensive program description. Such a need might arise, for example, in developing a manual to allow the replication of a substance abuse curriculum or curriculum element. Systematic program descriptions are a major vehicle for communicating with the field, and, as discussed later, a comprehensive program description is a prerequisite for any evaluation activities designed to assess program effectiveness. In all cases, the need is to answer the question:

What does this program look like as it actually operates?

Many of the evaluation activities undertaken for the purpose of program description are similar to those undertaken for the purposes of management and monitoring. The major differences between evaluation for the purpose of management and monitoring and evaluation for the purpose of description concern the time-frame of the data collection and the depth of the data collection.

As suggested, management or monitoring systems are ongoing and provide a dynamic picture of changes over time. By contrast, a descriptive evaluation usually attempts to capture a "still picture" of a program or program element as it operates at a given point in time. Of course, a program description may contain historical or developmental elements, but the emphasis is on the program as it operates now.

Descriptive evaluations strive for a greater depth than can be captured in the numbers generated by a management information system. For example, a management information system might reveal that in a given semester, X number of learners attended a newly offered selective course on anticipatory guidance for adolescent and pre-adoles-

cent patients and their parents. In addition, a descriptive evaluation would seek information concerning the learners' experiences in the course, how they reacted, and whether or not learners felt they could or would apply the course content in clinical practice. Gathering such information might require observation of course sessions, interviews with learners, debriefing of instructors, and so on. Such in-depth data collection strategies are the hallmark of descriptive evaluations. It is the fine-grained detail they provide that often makes them so useful in management decision-making and in enhancing assessments of program effectiveness.

Evaluation for Program Improvement

All professional schools are eager to improve the instructional program offered to learners. Program evaluation provides a mechanism for developing the information needed to make such improvements in a rational and planful manner.

Program improvement can take many forms. In some cases, it might mean attracting more learners to important courses or finding more appropriate clinical settings for clinical placements. In other cases it might mean making changes in the policies of the department or institution. In still other cases, program improvement might mean more effective or efficient learning, or increased impact on clinical practice behavior.

The questions that might be raised relevant to program improvement are numerous, but they all take the general form:

How can this program or program element be made more effective?

Sometimes the answers to these questions derive from management or monitoring systems or descriptive evaluations. Sometimes they require complex experimental designs. In either case, the emphasis is on identifying strengths upon which to capitalize and weaknesses in need of correction. In this sense, program evaluations for program improvement tend to fit with the traditional "valuative" connotations of program evaluation discussed earlier. However, the emphasis is on the use of data internally and in the overall service of the program.

Evaluation for Accountability

Evaluation for the purpose of accountability -- that is, evaluation for the purpose of determining whether or not a program is meeting its stated objectives -- also informs decision making. Often, however, the decision maker is someone outside the program (e.g., funders or university administration), and sometimes the decision is one of program continuation or termination. Thus, accountability evaluations can be threatening to those involved.

Accountability evaluations are similar to the evaluation activities described thus far, and may address many of the same questions. For example, an accountability evaluation of a clinical training program might address issues of course availability, course content, and course attendance (management or monitoring evaluation), learner experiences with and reactions to the courses (descriptive evaluation), and/or strengths and weaknesses (program improvement evaluation). Thus, a program that has these evaluative elements in place is generally in a good position to respond to requests for at least some kinds of accountability information.

Simply demonstrating an evaluation-based approach to program development, implementation, and improvement is one indication that a program is well run and rationally managed. Thus, programs in which evaluation is an ongoing and integral component of program operations tend to fare well when required to be accountable to outside decision makers.

Evaluation for Creating New Knowledge

A final purpose of evaluation is to develop and/or test hypotheses about programs and their effects. In the broadest sense, the goal of such evaluations is to contribute to the general knowledge base

about what does or does not lead to improvements in professional education, clinical practice behavior, and/or in patient outcomes regarding substance abuse.

If they are to be accepted by the scientific community, evaluations that seek to contribute to the general knowledge base must generally meet rigorous standards of research design. Many of these design issues are discussed in Section II of this *Guide Book*. It is important to note, however, that even the most rigorous evaluation design is greatly enhanced if other evaluative elements are in place (good program monitoring, a capacity for detailed program description, etc.).

It is also important to note that evaluations designed to produce new knowledge can have direct, practical implications for the institutions that conduct them. For example, an institution considering the adoption of interactive laser disc technology as a pedagogical strategy, or the use of standardized patients (actors trained to present a standardized complaint) as an assessment strategy may wish to compare these strategies to traditional, but less expensive strategies to determine if the improvement in outcomes is worth the extra cost.

MAJOR OPTIONS

Once the purpose of a given evaluation initiative has been determined, the evaluator must choose among available evaluation options. Many of these options are methodological and are discussed in detail in Section II. However, there are conceptual options that should be addressed before methodological options are considered. These conceptual options determine the overall direction and scope of the evaluation. They also guide, to some degree, the methodological decisions to follow. It is these conceptual options that are discussed here.

Program evaluation activities may be seen as varying along three major conceptual dimensions.

The overall goal of the evaluation (Formative, Summative)

The level of information gathered (Process, Outcome, Impact)

The type of information gathered (Quantitative, Qualitative).

Before discussing these dimensions in more detail, it is important to emphasize that they do not represent "either/or" options. For example, while an evaluation may be largely formative, it may have summative components. Similarly, a largely quantitative data collection effort may include qualitative data collection as well. Thus, evaluation activities are most often combinations of several components from each dimension, depending upon the overall purpose of the evaluation initiative.

The Formative/Summative Dimension

The terms "summative" and "formative" refer to the uses to which evaluation information is put. Formative evaluation information is used for program improvement and is usually reviewed on a regular basis as it is collected. A management or monitoring evaluation system in which data are reviewed each quarter or semester would be a relevant example. Summative evaluation information is most often used to assess the efficiency or effectiveness of a program that has reached a stable operational state. In contrast to formative information, summative evaluation information is often not analyzed until all data collection is completed in order to avoid contamination of the program under consideration. A controlled experimental study comparing an innovative curriculum module to a standard module is an example of a summative evaluation.

There is a tendency to associate formative evaluation with program descriptive information and summative evaluation with assessment of program outcomes. In fact, any type of evaluation may be either summative or formative, depending on the purpose of the evaluation and the uses to which the information is put. For example, a descriptive evaluation implemented in order to develop a

dissemination manual for an innovative program component (e.g., the use of standardized patients in a training) would be a summative evaluation because the purpose of the evaluation is to provide a finalized description of the component, and because the program component is stable at the point at which it is described. Similarly, an evaluation which provides an assessment of the effectiveness of a promising new educational technique would be considered formative if the purpose of the study were to aid in the further development of the technique.

In general, summative evaluation, whether oriented to description or outcome assessment will require a higher level of scientific rigor than will formative evaluation, and will, therefore, be more difficult and costly to implement. Keep in mind that it is inappropriate to carry out a summative evaluation when a program is still in a state of evolution, and when program components have not yet reached the level of stability required for a summative assessment.

The Process/Outcome/Impact Dimension

Some evaluators have found it useful to distinguish three levels of evaluation: process, outcome, and impact.

The *process* level is concerned with the quality and quantity of program inputs and activities, and with the socio-political context in which the program operates. Examples of the kinds of information subsumed by the process level include faculty and learner demographics, course descriptions, learner evaluations of courses, and context variables such as institutional support for curriculum changes.

Process data are the basis for program management and monitoring systems and for evaluations designed for the purpose of program description. Process data also play several key roles in evaluations designed to assess program effectiveness. At the simplest level, process data allow a determination of the *actual* program that is being evaluated. Programs as implemented are almost always different from programs as planned, and it is the effectiveness of the program as implemented that is of primary interest.

Process data also aid in unraveling the usually complex outcomes of programs and program components. For example, answers to the question "What pedagogical approaches work best with different categories of learners (e.g., undergraduates, post-graduates, fellows, faculty, etc.)?" are derived from an examination of differential effectiveness data in light of process information.

The preceding discussion suggests a general rule: Assessments of program effectiveness *should not be undertaken* in the absence of a well developed process evaluation.

The *outcome* level subsumes assessments of the accomplishment of specific program objectives concerned with changes in knowledge, attitudes, clinical behavior, patient outcomes, and so on. This simple conceptual definition belies the technical complexities involved in implementing adequate evaluations at the outcome level.

The technical complexities associated with outcome assessments derive from the necessity of meeting two challenges. First, measurement techniques must be located or developed that provide adequate representation of the outcome objectives of the program. Because substance abuse is a relatively new field in professional education, few standardized measures are available. Second, an evaluation design must be developed that supports the inference that it is the program or program component under study that has caused changes in attitudes, knowledge, clinical behavior, or patient outcomes -- that is, a design that rules out alternative explanations of the observed changes. A large proportion of the methodological discussions in Section II are concerned with strategies for meeting these two challenges of outcome assessments.

Evaluation at the *impact* level assess the broader effects of a program or the aggregate effects of several programs operating over time. Examining

the extent to which substance abuse training is institutionalized over a period of years would be one example of an impact evaluation. Impact evaluations also look for unintended program effects, both positive and negative. For example, a well regarded substance abuse training program in one professional school might stimulate the development of similar programs in other professional schools in the university. Generally, such unintended effects are discovered through descriptive data collection, and only the most complex and costly impact evaluations seek to study such effects systematically.

The Quantitative/Qualitative Dimension

The distinction between quantitative and qualitative evaluation is often made in terms of the *methods* used to collect data. Thus, knowledge tests, closed-ended interviews, and counts of patients referred for treatment are considered quantitative methods, while observation, open-ended interviews, and case histories are considered qualitative methods. The former are characterized as systematic and quantifiable while the latter are considered less systematic and less quantifiable. This distinction rapidly breaks down, however, as it becomes clear that almost any data collection can be made systematic and all data can be quantified. For example, observational methods may use standardized protocols and sophisticated time samples that yield precise data; open-ended interviews may be coded to produce reliable categories, and so on.

A more useful distinction is to consider the general paradigm that underlies quantitative and qualitative analyses. Quantitative analyses derive from a hypothetico-deductive paradigm in which a hypothesis or series of hypotheses are tested by relating variations in independent variables to variation in dependent variables. Such analyses require a substantial body of knowledge about the phenomenon (program) under study and the mechanisms that are thought to underlie its operation and/or effects. Here, precise definitions of variables and precision in measurement are key to the hypothesis testing process.

Qualitative analyses derive from a holistic-inductive paradigm which works backwards from the phenomenon (program) under study in an attempt to develop hypotheses about its operation and effects. Although data collection may still be highly systematic, the evaluator casts a much wider net than in quantitative analyses, attempting to establish patterns and regularities on which to develop insights into the mechanisms underlying operations and effects.

The choice of options along the qualitative/quantitative dimension will depend upon the level of existing knowledge concerning the program or program component under consideration and the nature of the questions being asked. Where there is sufficient knowledge to develop specific hypotheses and to define relevant variables precisely, a quantitative approach is indicated. By contrast, when relatively little is known about a program or component (as might be the case in a newly initiated, novel pedagogical practice), a more qualitative approach will generally provide a richer yield of useful information.

Frequently, evaluations evolve from qualitative to quantitative. For example, a given department might be interested in exploring the mentoring component of a faculty development program. Very little may be known at first about how the component operates, its relevant features, even its feasibility. The evaluation might start, therefore, by exploring the component qualitatively through semi-structured interviews with participating faculty, their mentors, and other key informants in the department (e.g., administrative personnel). Once the specific variables of interest have been identified qualitatively, they might be measured quantitatively through a closed-ended survey with quantitative rating scales.

Most successful evaluation efforts comprise a blend of both quantitative and qualitative data. For example, a study of learner attitudes towards substance abusers might rely on a paper-and-pencil assessment amenable to quantitative analysis supplemented by in-depth, open-ended debriefing of a

smaller sample of learners. Here, the more precise, but necessarily narrow, data from the quantitative analysis are supplemented by the descriptive depth offered by the qualitative analysis.

Quantitative data supplemented by qualitative anecdotes are also a powerful tool for communicating evaluation results. Evaluation practitioners have long known that, although quantitative rigor is the *sin qua non* of program evaluation, most individuals are more compelled by anecdotes and case examples than they are by statistical graphs and tables. This is not to suggest that evaluations should therefore be based solely on anecdotes and case examples. Rather, the point is that such qualitative data are useful communication tools when they are consistent with and support quantitative findings.

BENEFITS AND RISKS

Benefits

Many of the benefits of evaluation have already been described or alluded to in the preceding discussions. As suggested, the primary benefit of evaluation is that it provides a rational basis for program decision making -- including decisions concerning the most effective ways for integrating substance abuse content into professional education -- by providing systematic input into the decision making process. To this may be added at least four other benefits:

Evaluation as an early warning system

Evaluation as positive reinforcement

Evaluation as burnout prevention

Evaluation as legitimization

As already suggested, evaluation activities can provide regular feedback on the operation of a program or program component. One of the most important functions of such activities is their ability to serve as an *early warning system*, uncovering problems as they arise. A regular review of management or monitoring system data allows the identification of problem areas which might otherwise go unnoticed, and to take corrective action before these problems become serious.

Everyone is interested in feedback about job performance, and, in professional education, faculty are always interested in the impact their work is having on learners. Similarly, clinicians are extremely concerned about the impact they are having on the health and well being of patients or clients. An evaluation system can be of direct benefit in meeting these needs and can provide *positive reinforcement* to individuals who might otherwise have little systematic way of assessing their own performance.

Related to the reinforcing value of evaluation is its role in *preventing "burnout."* It has often been observed that the introduction of evaluation activities into a program provides a new and interesting focus for individuals who may be somewhat jaded or discouraged. It is also the case that the self-examination that accompanies the introduction of evaluation activities into a program provides a positive forum for discussion of operational issues that might otherwise remain hidden. Repeated observation of this self-examination phenomenon has led some evaluators to refer to evaluation planning activities as the "evaluation therapy" process.

Educating the public and other important constituents about programs -- i.e., *public relations* -- another important function of evaluation data. As noted, descriptive evaluations are often used as communication tools to answer the common question, "What exactly does this program do?" Similarly, management or monitoring systems allow ready answers to the four journalistic "W's" (who, what, where, and when), issues often raised by outside individuals interested in program operations.

Finally, the *legitimizing* function of evaluation must be considered. All else being equal, a pro-

gram that is actively engaged in examining its activities and its impact will be viewed as more legitimate than a program that is not. The legitimacy that derives from an active program of evaluation is well warranted -- programs that include evaluation as part of their ongoing activities are also more likely to be those programs that are effectively managed and whose growth and development are rational and planful.

Risks

The most commonly discussed risk of evaluation is that the data will reveal something negative about the program or its impact. This perspective derives from the view of evaluation as largely a "valuative" activity. As has been repeatedly emphasized, the "valuative" function of evaluation is by no means its only, or even its primary function. Even when evaluations *are* designed to directly assess program worth or efficacy, evaluation results are almost always a mixture of positives and negatives. Rarely, if ever, do evaluation results deliver the programmatic *coup de gras*.

This is not to imply that evaluation activities are without risks. However, these risks tend to have little to do with what the results of the evaluation reveal about the program. Rather, they involve the practicalities of introducing evaluation activities into the ongoing operations of a program. Specifically:

Evaluations can be threatening to those who have an investment in program success.

Evaluation activities can interfere with other activities.

Evaluations can take away resources from program activities.

Despite all arguments to the contrary (including those presented in this manual), some people involved with a program will be initially *threatened* by the prospect of initiating evaluation activities. Even when it is understood that the evaluation will not be used to make judgments about the adequacy of performance (i.e., it is the program, not the people, that is being evaluated), evaluation can still be threatening because evaluation may portend change and a threat to the *status quo*. In addition, many people harbor the suspicion (sometimes justified by previous experience with evaluators), that the tools of evaluation are ill-suited to measure the quality of certain program activities or to assess "intangible" outcomes.

Experience suggests that the threat of evaluation can be reduced by including the individuals whose program is to be evaluated in the evaluation planning process and by providing them with an active role in the interpretation of evaluation results. Moreover, certain evaluation components (e.g., management or monitoring systems) can be of benefit to faculty and other staff by providing regular feedback as discussed earlier.

Experience also suggests that autocratic attempts to impose evaluations on resistant participants will usually fail. The ways in which an unwanted evaluation can be undermined are too numerous to list -- it suffices to say that without full cooperation, most evaluation activities will produce little in the way of useful information.

Almost all evaluation activities obtrude, in some way, into ongoing program or institutional activities. Such *interference* can be minimized by careful selection of evaluation methods, but the fact remains that additional activities will be required that are not always consistent with programmatic objectives. For example, an evaluation design may include the observation of interactions with patients or clients, and it may be argued that the presence of the observer negatively affects the provider/patient relationship. Similarly, time devoted to an extensive knowledge or attitudes pre-test is time taken away from instruction. There is no "magic bullet" to resolve these conflicts between the evaluation agenda and the program agenda. Rather, a balance must be struck between compromise in the program and compromise in the evaluation design given the probable benefits of

implementing a given evaluation activity.

One final risk of evaluation bears consideration -- the risk of misusing evaluation results. In some instances, slipshod or equivocal evaluations are used to support claims of program effectiveness that are not justified by the evaluation design or by the data. Here, the loss of credibility can be sufficiently damaging to the program or institution as to overshadow any benefits that might derive from the "positive" evaluation results.

Similarly, design or implementation weaknesses in the program itself or in the evaluation may lead to negative evaluation findings that are not a true reflection of the effectiveness or potential effectiveness of the program. An otherwise sound program idea may be prematurely rejected.

Both promotion of an unsound program and the rejection of an effective one are serious risks -- they can best be avoided by using the most rigorous methodology feasible and by interpreting evaluation results carefully.

SECTION II:
INTRODUCTION TO RESEARCH METHODS

This section provides an overview of the methodological issues involved in evaluating faculty development and clinical training programs in substance abuse. The overview is intended to raise key issues in program evaluation research methods, and to relate these issues to the specific problems faced in the design and implementation of the various kinds of evaluation activities described earlier (monitoring and management systems, descriptive studies, accountability studies, etc.). Where appropriate, references are provided to more detailed, technical descriptions for exploration by the interested reader.

SAMPLING

Sampling involves strategies for selecting a portion of cases from or about which data will be collected with the goal of making generalizations about the larger population from which the cases are drawn. For example, in evaluations focused on learner outcomes, a case might be a single learner, the enrollees in a course, or all the learners in a department. Similarly, the population to which generalizations are to be made might be all learners enrolled in a given course, all learners in a given clinical specialty, all departments within a professional school, and so on.

In some instances, an evaluation design may call for data collection from or about *all* cases in a population. This might be the case, for example, in a management or monitoring system that tracks learner outcomes from year to year based on standardized test scores. More often, however, it is not practical, or even desirable, to collect data from an entire population. Under these circumstances, a sampling strategy is required.

Sampling strategies may be divided into two broad categories -- those that yield samples with known or measurable sampling errors, and those that do not. Sampling error is a measure of the precision with which the parameters of a sample represent the parameters in the larger population. From the standpoint of methodological rigor, samples with known sampling errors are clearly preferred. However, such samples may not always be practically or conceptually feasible, and other sampling strategies must, out of necessity, be employed.

All samples, with or without a known or measurable sampling error, may be further categorized according to their variability. In general, samples with lower variability are preferred, owing to the inverse relationship between variability and statistical power (i.e., the ability to detect effects if they are present). It is important to note that the preference for samples with lower variability applies

in *qualitative* studies may be quite different -- in qualitative studies, samples may be constructed in such a way as to *increase* variability in order to gather as diverse a data base as possible.

The following sections describe the major sampling strategies used in program evaluation. Strategies yielding samples with and without known sampling errors are discussed separately.

Samples with Known Sampling Errors[1]

Simple Random Sampling: In simple random sampling, each population member (case) is assigned a unique number. The sample is then selected via use of random numbers. Simple random sampling has three advantages: 1) it requires only minimum knowledge of the population *a priori*, 2) it avoids classification errors, and 3) it facilitates analysis of data and computation of errors. Disadvantages of simple random sampling include that it does make use of knowledge of the population which might be possessed by the evaluator, and it yields larger errors (for the same sample size) than does stratified sampling.

Systematic Sampling: Systematic sampling exploits the natural ordering of a population. A random starting point is selected between the number one and the nearest integer to the "sampling ratio" -- defined as the ratio (n/N) between the size of the sample (n) and the size of the population (N). Items are then selected at the interval nearest (at the whole number) to the sampling ratio. If the population is ordered with respect to some pertinent property (e.g., department, clinical specialization), then systematic sampling yields a stratification effect (e.g., insures roughly equal representation of departments or specialties). This stratification reduces variability compared to a simple random sample -- the major advantage of systematic sampling. In addition, systematic sampling facilitates both the drawing and checking of the sample.

However, if the sampling interval is related to a periodic ordering of the population (e.g., substance abuse-related emergency room admissions by time of day or day of the week), increased variability may be introduced. This is the major disadvantage of systematic sampling. When there is this type of stratification effect, estimates of error are likely to be high.

Multistage Random Sampling: Multistage random sampling involves stages, all of which are a form of random sampling. For example, in a follow-up study of patient outcomes, clinical sites might be randomly selected (stage one), and then patients or clients randomly selected within sites (stage two). A major advantage is that sampling lists, identification, and numbering are required only for units belonging to subgroups (e.g., clinical sites) actually selected. If sampling units are geographically dispersed (e.g., satellite clinics), multistage random sampling cuts down on the time and trouble needed to collect data.

On the negative side, errors are likely to be larger for multistage random sampling than for simple random or systematic sampling for the same sample size. Errors increase as the number of sampling units selected decreases -- if, for example, only a small number of satellite clinics can feasibly be sampled.

In a major variant of multistage random sampling, sampling units are selected with probability proportionate to their size. This procedure has the advantage of reducing variability. Its major disadvantage is that lack of *a priori* knowledge of the size of each sampling unit increases the variability.

Stratified Sampling: Stratified sampling strategies make use of knowledge about characteristics of the population (e.g., age, sex, and ability to pay of patients in a clinic) which may be related to dependent variables (e.g., patient referrals). There are three major variants: 1) proportionate sampling, 2) optimum allocation sampling, and 3) disproportionate sampling. Each of these three variations is discussed in turn.

In proportionate stratified sampling, selection from every sampling strata is random with probability

proportionate to size. This assures representation with respect to the property or properties which define the strata and, therefore, yields less variability than simple random sampling or multistage random sampling. Proportionate stratified sampling also decreases the chance of failure to include members of the population because the classification process helps insure that a wider range on the variables that define the strata will be included. In addition, the stratified sample facilitates comparisons between or among strata (e.g., adult patients vs. adolescent patients). On the negative side, proportionate stratified sampling requires accurate information on the proportion of the population in each stratum. Otherwise, error will be increased. If stratified lists are not available, these may be costly to prepare. There is also the possibility of faulty classification and hence, increased variability.

Optimum allocation sampling procedures are the same as those in proportionate sampling, except that the sample is proportionate to the variability within strata as well as to their size. This assures that there will be less variability for the same sample size than in proportionate stratified sampling. The major disadvantage is that optimum allocation requires knowledge of variability of pertinent characteristics within each stratum. For example, it may be useful to classify patients within clinics according to variability in the nature and extent of substance abuse problems. However, accurate assessments of such variability may be difficult to obtain.

Disproportionate stratified sampling proceeds as in the proportionate and optimum allocation variants, except that the size of the sample is not proportionate to the size of the sampling units, but is determined rather by analytic considerations or convenience. For example, one might choose to oversample a cell within a stratified sampling design if the proportion of this cell within the general population is so low as to yield unanalyzable results (for example, oversampling patients of a particular sex, age, and usage patterns such as cocaine addicted pregnant teenagers). Disproportionate stratified sampling is more efficient than proportionate stratified sampling for comparison of strata or where different errors are optimum for different strata. The major disadvantage is that disproportionate stratified sampling is less efficient than proportionate sampling for determining population characteristics. That is, it yields higher variability for the same sample size.

Cluster Sampling: In cluster sampling, sampling units are selected via some form of random procedures. The ultimate units are groups (e.g., a clinic, enrollees in a course). Cluster sampling has several advantages: 1) if clusters are geographically defined, it yield the lowest field costs, 2) it requires listing only units in selected clusters, 3) characteristics of clusters as well as those of the population can be estimated, and 4) it can be used for subsequent samples because clusters rather than units are selected, and substitution of units may be possible. The disadvantages are larger errors for comparable sample sizes than with other probability samples, and the requirement that each member of the population be uniquely assigned to a cluster -- inability to do so may result in duplication or omission of units.

Stratified Cluster Sampling: Stratified cluster sampling is to cluster sampling what stratified sampling is to simple random sampling. That is, sampling strata are defined based on known characteristics of the clusters, and clusters are selected at random from every sampling strata. Stratification reduces the variability of ordinary cluster sampling, but combines the disadvantages of stratified sampling with those of cluster sampling. In addition, because cluster properties may change, the advantage of stratification my be reduced, and the sample may not be usable for subsequent research.

Sequential Sampling: Sequential sampling is a procedure whereby two or more samples of any of the types discussed above are taken, with results from earlier samples used to design later ones (or to determine if they are necessary). Sequential sampling provides estimates of population charac-

teristics that facilitate efficient planning of succeeding samples, and thereby reduces errors in the final estimate. In the long run, sequential sampling also reduces the number of observations required. It has several disadvantages: 1) it complicates the administration of field work, 2) it requires more computation and analysis than does non-repetitive sampling, and 3) it can be used only where a very small sample can approximate representativeness and where the number of observations can be increased conveniently at any stage of the research.

Samples Without Known Sampling Errors

Convenience Sampling: The simplest sampling technique is convenience sampling. Here, the evaluator collects data on those individuals who are most readily available. An example would be an assessment of changes in learners' patient interviewing skills using learners assigned to a certain clinic chosen on the basis of geographic proximity (i.e., it is easy to get to) or timing (the clinic hours fit the schedule of the data collectors). Although the convenience sample is one of the weakest available sampling options, it is usually sufficient for pilot studies or for studies aimed at very general assessments.

Quota Sampling: Quota sampling attempts to overcome some of the disadvantages of convenience sampling by introducing stratification based on *a priori* knowledge of a population. For example, in a study of case notes entered by learners exposed to an educational innovation, clinic records might be used to determine the proportion of male and female patients in different age groups, and "quotas" set for male and female patients of different ages that represent the population proportions. Quota samples reduce variability and increase representativeness, although the selection of cases within categories (e.g., male, female) may still be highly biased.

Modal Sampling: Modal sampling involves the selection of a sample that is judged to be representative of a population based on knowledge of the general characteristics of the population of interest. For example, experience might suggest that patients in a given clinic tend to fall into categories characterized by a set of related characteristics (e.g., young males who are heavy drinkers but not alcoholics; older males who are either light drinkers or alcoholics, etc.). The modal sample is selected to include individuals from each of the various categories which, taken together, are thought to represent the range of patients in that population. Modal sampling is particularly useful in evaluations where a large amount of resources are to be devoted to the study of a small number of cases, as might apply when a study requires lengthy interviews with patients.

In general, the quality of a modal sample will depend upon the extent and sophistication of the evaluator's understanding of the characteristics of the population. Like quota sampling, modal sampling has the advantage of increasing representativeness and decreasing variability, and like quota sampling, the possibility of biased selection within modal categories is a major disadvantage.

Purposive Sampling: Purposive sampling is used in qualitative evaluation to ensure that information is obtained from each group or individual of interest. For example, in order to describe the changes that have occurred as the result of the introduction of a new curriculum, key informants (such as the department chair and some professors) might be selected from each participating department.

Snowball Sampling: Snowball sampling is a variation on purposive sampling. Key informants are selected and each of these informants is asked to suggest other people from whom data should be collected. For example, investigators interested in the opinions of students most interested in alcohol and other drug abuse issues might ask the professors of key courses to suggest students to be interviewed. Those students might in turn be asked to suggest other students who share their interest in alcohol and other drug abuse topics. This process could continue until a sample of sufficient size has been identified or until most of the suggested informants have already been named by others.

ISSUES RELATED TO SAMPLE SIZE

One of the most commonly asked questions in planning an evaluation is: "How large a sample is required?" Actually, there are two different questions related to sample size. The first question is: "How big a sample is required in order to make reasonable generalizations from the sample to the larger population?" The second question is: "How big a sample is needed in order to make meaningful inferences about effects?"

Sample Size and Generalizability

The first question relates to the issues of sampling error discussed earlier -- that is, how well does the sample represent the population. In general, the larger the sample size, the smaller the sampling error, and hence, the better the representation of the population of interest. Specific formulae for calculating sampling error for different sample types and strategies can be found in Kish (1967).[2] As in most things, a law of diminishing returns applies to sample sizes. As sample size increases, the relative increase in precision derived from each additional case *decreases*. Thus, evaluators sometimes refer to the "net gain from sampling," a comparison between the increased cost in terms of data collection, data analysis, etc., of each additional case and the concomitant benefit in terms of increased precision.

When sampling strategies are employed that yield samples for which sampling error cannot be measured, the sample size is largely a practical judgement based on the available resources for the evaluation, and a consensus among those who will use the evaluation results. Here, the issue becomes one of negotiating the number of cases needed to instill confidence in those individuals who will be using the information the evaluation produces.

Sample Size and Statistical Power

The second question relates to statistical power, that is, the ability of inferential statistics to detect effects when such effects exist. The power of a statistical test represents a complex relationship among three factors:

The probability of mistakenly rejecting the hypothesis that there is no effect when, in fact, no effect exists (referred to as alpha) -- as alpha increases, power increases

The magnitude of the true effect size in the population -- as the magnitude of the effect increases, power increases

The sample size -- as the sample size increases, power increases.

Knowledge of two of these three parameters allows a calculation of the third. Alpha is set by the evaluator. Thus, knowledge of the true effect size allows a calculation of the sample size needed to obtain a given level of power. Conversely, for a given sample size, the minimum detectable effect can be estimated. Procedures for making these calculations are found in Cohen (1969) and Cohen and Cohen (1975).[3]

Unfortunately, the true effect size is generally not known. Cohen and Cohen (1975)[4] suggest three strategies which allow a calculation of sample size in the absence of *a priori* knowledge of the true effect size. First, other similar studies may be consulted in order to ascertain the probable effect size in a given situation. For example, in order to estimate the expected change in learner knowledge scores as a result of participation in a given course, one might examine changes observed in other studies of similar courses.

Second, one might consider the size of effect that would make a practical difference -- e.g., how big a change in clinical practice behavior would be required to make it worth implementing a given strategy to improve a given clinical skill?

Finally, Cohen (1969)[5] has suggested conventions for "small," "medium," and "large" effect sizes based on the effects generally observed in the social sciences. In the absence of other informa-

tion, sample sizes may be calculated for all three "conventional" effect sizes in order to establish a range of sample sizes for a given evaluation study.

THE VALIDITY OF EVALUATIONS AND POTENTIAL SOURCES OF BIAS

The overall goal of the various social science methods used in evaluation is simply to ensure, to the greatest extent possible, that the information produced is accurate enough to yield valid conclusions useful in decision making. Stated another way, the goal of these methods is to reduce, to the greatest extent possible, the *biases* which may affect the *validity* of evaluation conclusions.

Issues Related to External Validity

All evaluation seek to maximize *external validity* -- that is, the extent to which the results of a given evaluation are generalizable beyond the highly specific conditions under which the results are produced. External validity will be of concern whether the evaluation is a management or monitoring system, a descriptive evaluation, or a complex assessment of program outcomes or impact. In general, the biases that impair external validity are to be found in the ways that samples are drawn and measurement is conducted.

Sampling Biases: Several issues related to sampling have already been discussed. As suggested, various sampling techniques yield samples which estimate population parameters with varying degrees of precision, and hence, will yield evaluation results with varying degrees of external validity. When samples are not systematic, and often they are not, biases will be introduced that decrease the meaning of results beyond the specific population from whom data are gathered. At the very least, these biases should be acknowledged in reporting evaluation results -- better still, attempts should be made to estimate and/or correct for their effects.

Two of the most common sample biases that arise in program evaluation -- and these obtain no matter how the initial sample is drawn -- are *refusal bias* and *attrition bias*. *Refusal biases* may occur whenever a study population is given the option to refuse to participate in some or all of the data collection activities. The ethics of evaluation (and the rules governing all Federally funded research) require that subjects be given the option to refuse cooperation, to terminate cooperation at any time, and to refuse to answer individual questions. Although careful attention to the rights and concerns of study participants can reduce refusal, some refusal is inevitable in most evaluation data collection. For example, a common example of refusal bias is non-return of mail surveys. The question then arises: "How are subjects that refused to provide data different from those who did not refuse?". That is, what biases are introduced by refusal?

There are a number of important ways in which those who refuse to cooperate may differ from those who do not. For example, it might be assumed that the patients most likely to refuse to participate in a survey of the incidence of alcohol and other drug problems in a given clinic population are those with the most serious alcohol or other drug problems. Estimates of problem rates will be biased if those with the worst problems refuse to cooperate. A completely different sort of bias might be introduced if those with the least education refuse to cooperate to avoid the embarrassment of admitting that they cannot read or do not understand the data collection instrument.

Refusal biases may also arise in system level studies. For example, one or more departments might refuse to participate in a study of departmental support for the introduction of substance abuse-related curricular content. Is such a refusal the result of a desire to prevent the discovery of existing problems? Alternately, does the department chair believe that the department is already doing an adequate job in this area and that no benefit will derive to the department from the study? In either case, the study would be biased by the lack of data from the non-cooperative department(s).

Related to refusal biases are *attrition biases* --

those biases that arise when study participants do not complete the study. Like refusal, attrition raises the concern that those who do not complete the study may be systematically different from those that do. So, for example, a learner satisfaction survey taken at the end of a course may show inflated satisfaction because it does not include those highly dissatisfied individuals who dropped out. Similarly, a comparison might be made of lifestyle changes observed in patients counseled by clinicians who have or have not been exposed to an educational innovation. This comparison will be biased if attrition of heavy drug or alcohol users is differential between the two clinician groups.

Assessment of the impact of refusal or attrition biases generally relies on comparisons of the characteristics of cases from which complete data are available and those cases from which they are not. Of course, such a strategy requires some minimum level of data on lost participants. Thus, even patients who refuse to complete a questionnaire or interview might be asked if they would be willing to provide basic demographic information. Sometimes, assessment of the effects of refusal or attrition biases must be based on very rudimentary data such as race, sex, and approximate age. Such data might be gathered from case records or even through observation.

Measurement Biases: Accurate generalizations are possible only to the extent that accurate measures are used in the evaluation. In general, the external validity of evaluation studies is threatened when measures either fail to accurately represent the phenomenon of interest (so-called failures of *construct validity*), or systematically over- or underestimate the phenomenon of interest. Of course, measures may also be "noisy" or unreliable. However, low reliability is generally considered to be a threat to internal, rather than external validity.

The issue of construct validity is a major concern in evaluations of substance abuse programming. How does one measure (or even define) such key constructs as alcoholism, addiction, dependency, denial, social drinking, and so on. To say that a given intervention is successful in identifying "social drinkers" but not alcoholics requires a consensus about what is meant by these terms, and a consensus that the measurement techniques used to assess them are adequate representations .

Even as simple a construct as whether or not a learner completed a course raises definitional and measurement issues. Does the learner have to attend all sessions? If not, how many may he or she miss? Does the learner need to do all the required reading in order to be categorized as a course completer?

Assuming that a satisfactory definition of a given variable is available, and assuming that one can adequately measure the variable in some practical way, concern may still be raised abut systematic over- or underestimation. For example, if a learner's diagnostic classifications are being compared to the results of a written assessment tool, does the assessment tool over- or underestimate the number of patients with a given diagnosis?

Issues Related to Internal Validity

Internal validity refers to the extent to which a given evaluation design allows one to rule out alternative explanations for the effects observed. In general, then, internal validity is of concern when one is attempting to derive causal relationships among variables (e.g., participation in a given practicum *caused* improvements in clinical skills). The various threats to internal validity may be defined in terms of various challenges to the assertion that "exposure to X caused effect Y." Discussions of the common threats to internal validity follow.[6]

History: Historical threats to internal validity refer to specific events or conditions -- in addition to the treatment or experimental influence -- which occur between first and subsequent measurements of the dependent or outcome variable, and which might influence the magnitude of observed differences.

For example, an observed change in attitudes regarding alcohol and other drug abusers might be taken as evidence of the success of an educational intervention. However, such changes might actually have resulted from a widely publicized incident such as the overdose death of a famous athlete or a major oil spill attributed to alcohol abuse. Similarly, increased identification of drug abusers in the emergency room might result from the increased diagnostic skills of E.R. staff, but might also be attributable to actual increases in the number of drug abusers who are present at the E.R. during the study period.

Maturation: Maturation refers to changes in study participants between pretest and subsequent testing which influence the observed outcome, but are not a part of the program or treatment of interest. Participants may grow older, wiser, stronger, or undergo biological change.

For example, it is well known that drug and alcohol use generally increases throughout the teenage years, peaks in early adulthood, and then declines. Thus, depending upon the age of patients in a study, use rates may either increase or decrease over time independent of the exposure to any clinical intervention. Similarly, most psychopathology (including alcohol and other drug problems) is associated with a "spontaneous recovery rate" -- i.e., some individuals will improve in the absence of any treatment.

Testing: Sometimes, simply taking a test affects participant's subsequent performance on the same test. In pretest/post-test designs, the concern is with the influence of the pretest experience on the post-test score. For example, pretesting learners' attitudes concerning alcohol and other drug abusers may sensitize learners to the fact that attitude change is desired. The potential influence of testing is ever more likely in time series or repeated measurements designs where multiple testing is required.

Instrumentation: Instrumentation artifacts arise when changes occur in measurement instruments or procedures over repeated observations. For example, measures derived from observations of clinician-patient encounters may change in some unknown way as the study progresses and repeated observations are made. Observers may become more skilled, bored, or inferential as they rate samples of clinician behavior. Similarly, if measures are based on interview data, the quality of the data may change as interviewers gain experience and confidence, become more familiar with the interview schedule, or gain insight into the content of the interviews.

"Objective" data sources are also subject to instrumentation artifacts. For example, patient record keeping systems may be altered or improved, new or refined diagnostic categories might be added to standardized nosologies, laboratory tests may become more sensitive, and so on.

Statistical Regression: Statistical regression artifacts may arise when study participants are classified into or selected from extreme groups on the basis of pretest scores, correlates of pretest scores, or some other basis. When participants are assigned to groups on the basis of high or low pretest scores, the high groups will tend to score lower, and the lower groups higher at the post-test. For example, if patients are assigned to a specialized intervention based on the severity of their alcohol or other drug problems, these individuals will tend to exhibit fewer problems in the future due simply to statistical regression. Similarly, the often observed effect in psychotherapy that very disturbed individuals tend to improve more than less disturbed individuals is probably due to statistical regression effects rather than to a relationship between the extent of psychopathology and the effectiveness of psychotherapy.

Statistical regression effects are especially likely if the pre- and post-tests are unreliable, or include a considerable amount of measurement error. Under these conditions, changes from pre- to post-test are very likely due to statistical regression, and attributing such changes to program influences would almost certainly be incorrect.

Selection: As used here, selection refers to a treatment effect due to a lack of initial equivalence between study groups. For example, the effects of an innovative pedagogical strategy might be tested by recruiting volunteers for the new strategy from all learners enrolled in a particular course. Under these conditions, differences observed between the two groups of learners might be due to initial differences between those who volunteer and those who do not, rather than to any particular benefit of the new strategy.

Mortality: Mortality artifacts arise when participants drop out as the evaluation progresses from inception to completion. Here, differences between pretest assessments and subsequent assessments may reflect changes in the study population rather than any effects of the intervention.

In comparisons among strategies, there is always a concern that mortality will be differential -- that is, different types of participants will be lost in the different treatment conditions. For example, a comparison might be made between two methods for improving learner attitudes towards abusers. The first might involve a series of lectures, while the second might include lectures plus guided discussion of the learners' personal experiences with alcohol and other drug abuse and with abusers. The latter strategy might cause learners with the most negative attitudes or most serious personal problems sufficient discomfort that they drop out of the course. Thus, differential changes in attitudes in the two groups would be observed, but these changes could not reasonably be attributed to differences between the two educational methods.

Interactions with Selection: When cases are not randomly assigned to conditions, the possibility exists that certain characteristics of study groups will interact with other threats to internal validity and produce additional spurious treatment effects. For example, the influence of maturation on a treatment group might be different from that on a comparison group if the treatment and comparison groups are not initially equivalent. This might be the case for example, if patients exposed to a treatment are drawn from two clinics that differ in their age distributions.

Ambiguity About the Direction of Causal Influence: The direction of causal influence is a threat to internal validity in correlational studies, especially when an equally plausible argument can be developed for the conclusion that A causes B, or B causes A. For example, a study might assess the association between learner satisfaction with a course and the degree to which the learners apply the course content in clinical practice. One might conclude from such a study that learners who liked the course are more likely to apply it than learners who do not. Equally plausible, however, is the conclusion that learners who apply the content are more likely to report that they liked the course than are learners who do not apply the content.

COMPARISON AND CONTROL GROUPS

Imagine the following scenario:

Learners entering an innovative clinical practicuum are given a battery of assessments, including alcohol and other drug knowledge and attitudes, accuracy in diagnosing dependency, and frequency of referrals of patients to treatment. Following the practicum, the assessment battery is repeated. Comparisons of the test scores prior to and after the practicuum reveal that knowledge and attitudes improved, diagnostic accuracy remained about the same, and frequency of referrals went down.

What can be concluded about the effects of the practicum from the results obtained? Based on the material discussed thus far, the answer is: "Not much!"

The reason that not much can be concluded from this hypothetical study is that there is no basis for ruling out a variety of other explanations of the observed results -- that is, no way of assessing threats to internal validity that may have caused the observed changes (or lack of change) in the absence of any program effects. For example, the

observed improvement in knowledge and attitudes might be a result of the learners' exposure to articles appearing in widely read journals, while the observed reductions in referrals might result from the fact that a major treatment facility stopped accepting new patients (both are historical artifacts). Even the *lack of change* in diagnostic accuracy is suspect -- perhaps diagnostic accuracy would have gotten worse without exposure to the practicum as learners were assigned increasingly more difficult patients (an instrumentation artifact).

The single major solution to the problems raised by threats to internal validity is the use of control or comparison groups -- that is, the testing of individuals who are identical to the treatment population in every way, *except* for exposure to the treatment. The results obtained for the control or comparison group are those which are attributable to history, maturation, testing effects, etc. The residual differences observed in the treatment group are those attributable to the treatment. So, for example, if a control group were used in the study discussed, and if these individuals showed *decreased* diagnostic accuracy over the study period, the practicum would be deemed successful by virtue of having stemmed this rate of decrease.

The difference between a control and a comparison group is that control groups are developed by *randomly* assigning cases to treatment, while comparison groups are developed in any other way. The great majority of evaluation studies rely on comparison groups rather than randomly assigned controls. Unfortunately, the ability of comparison groups to rule out certain threats to internal validity is often weak, and statistical methods for adjusting initial group differences, although commonly employed, can never approximate the inferential power that derives from true control groups. On the other hand, a non-randomly created comparison group is much better than no comparison at all, and may be the only practical option in many evaluations.

A major roadblock to the use of either control or comparison groups is the fact that, for some individuals, the opportunity to participate in a given program (i.e., the one under study) must be withheld. In studies of patient outcomes, withholding a treatment seen as potentially beneficial may cause ethical or legal problems. However, a number of options exist for constructing control and/or comparison groups which may be acceptable and feasible within a given setting. These are discussed below.

The term "control group" is used throughout the discussion that follows to refer to both control (randomly created) and comparison (non-randomly created) groups. Where no distinction is made, either a random or non-random assignment strategy is possible. -- i.e., either a control or comparison group is possible. In cases where only a non-random assignment strategy is possible, the term "comparison group" is used.

The True No-Treatment Control

In some evaluations, the most desirable control condition is one in which some cases simply do not receive the treatment under study. For example, participants in an innovative substance abuse practicum might be randomly selected from among those learners who express an interest in participating. Those not selected to participate serve as the control group. Such a strategy is often justifiable on practical grounds when the number of available slots exceeds the number of interested learners.

The Waiting List Control

The waiting list control is implemented by assigning some cases to receive the program now and some cases to receive the program later. All cases are pretested before the "now" group receives the program and are post-tested after the "now" group completes the program. The "later" group thus serves as a non-treated control during the waiting period. As is the case with the no-treatment control, the waiting list control provides a practical solution when demand for a program exceeds supply. The waiting list control has the disadvan-

tage that the waiting list may not truly provide a "no treatment" comparison in that individuals may receive some "stop gap" services. In addition, it precludes any long-term comparisons. As soon as those on the waiting list enter the program, all comparisons between the groups are nullified.

The "Traditional Treatment" Control

The "traditional treatment" control compares an innovative strategy to an existing strategy (traditional treatment). For example, an innovative training program might be compared to an existing training program. Although this design does not allow an assessment of the *absolute* effects of either the "traditional" or the innovative strategy, an assessment of the relative effectiveness is available.

The Minimal Treatment Control

The minimal treatment control is one in which minimally acceptable treatment (defined by law, licensing guidelines, program standards, or ethical considerations) is used as a comparison. For example, in-depth anticipatory guidance for parents of adolescents could be compared to a condition in which parent education pamphlets are simply available in a clinic waiting room.

The Placebo Control

When a control treatment is so minimal that it is expected to have no effect whatever, it is referred to as a placebo control. Sometimes, placebo controls are used to determine the potential operation of so-called "Hawthorne Effects" -- that is, the effects of simply paying attention to study participants.

The Non-Eligible Comparison Group

Some evaluation studies use individuals who are not, for one reason or another, eligible to receive the treatment as a comparison group. The term "comparison group" is used here because assignment to such a group is necessarily non-random. For example, learners whose schedules preclude them from taking an elective course on patient interviewing might constitute a group to which learners who take the course may be compared. This strategy presents a number of conceptual problems since the criteria that determine eligibility or non-eligibility may be related in a variety of ways to the outcomes of interest.

MEASURING OUTCOMES

Faculty development and clinical training programs seek to attain a variety of outcomes, including improvements in knowledge, attitudes, and clinical skills, organizational change, and improvements in patient outcomes. The specific outcomes measured in an evaluation of any aspect of a faculty development or clinical training program will be determined by three factors: 1) the specific objectives of the program, 2) the questions the evaluation is designed to answer, and 3) the scope and time frame of the evaluation effort.

The relationship between program objectives and program outcome measures cannot be emphasized too strongly. Perhaps the most common error in all outcome evaluations is the use of measures that do not relate directly to what the program is attempting to accomplish. For example, improved awareness of learners' own use of alcohol and other drugs is a worthwhile goal, and the evaluator may be motivated to measure changes in this area as part of the outcome evaluation of an educational innovation. But if the program is not *specifically* designed to increase awareness (for example, if the innovation is designed to teach pharmacology), the choice to measure awareness may result in an artificial finding of program ineffectiveness -- i.e., awareness does not increase because there is no particular reason why it should.

Evaluators must also consider the use to which outcome information is to be put. In some cases, such as a summative evaluation of a well established program or program component, highly

precise (and probably expensive) measures of outcome may be required. Indeed, as discussed below, multiple, independent outcome measures would be appropriate in such an evaluation. In other cases, however, less precise (and less expensive) measures may be sufficient. For example, a preliminary assessment of the effectiveness of an innovative educational strategy might be accomplished by asking for student feedback about whether the innovation seemed useful and appropriate. Such measures may be all that is required to evaluate whether or not the program element is worth developing further.

Finally, the scope and the time frame of the evaluation must be considered. For example, long-term patient outcomes might be a very important indicator of the success of an effort to improve clinical skills. However, such an assessment may require studying large numbers of patients or following patients for significant periods of time. This is especially the case if the incidence of alcohol and other drug problems in the patient population is low -- e.g., adolescents or pre-adolescents.

: is also the case that many outcome measures .e.g., the percentage of patients for whom alcohol and drug histories are taken) might show significant changes immediately following exposure to a seminar on history taking. However, these changes may decay within a short period of time. Therefore, if a reasonable follow-up period cannot be accommodated, the evaluator may choose to deemphasize such measures in order to avoid basing program decisions on highly unstable outcomes.

Generally, the outcomes of faculty development and clinical training programs are measured in one of four ways: 1) self-reports, 2) other-reports, 3) observation of behavior, and 4) review of records. The strengths and weaknesses of each of these measurement techniques follow.

Before proceeding to a discussion of each of these types of outcome measures, it is important to note that the most conclusive evaluations of outcome *rarely* rely on a single type of outcome measure. This is because all outcome measures have weaknesses that limit the strength of conclusions based on them. Thus, when multiple measures are employed, the strengths of one measure may balance the weaknesses of another. Moreover, to the extent that the various measures coincide (i.e., suggest the same conclusion), the inferential strength of the evaluation increases. The use multiple measures is an example of "triangulation" in social science research.[7]

Self-Report Measures

Self-report measures include those in which respondents are asked to assess the quantity or quality of their own behavior, or to provide a variety of other information that may be used to assess knowledge, attitudes, demographics, and so on. The usual methods for gathering self-report data are paper-and-pencil questionnaires and interviews.

The major advantages of self-report data are their ease of collection and relatively low cost, and their generally high *reliability* -- i.e., freedom from internal measurement error. In addition, self-reports are the only practical way to assess certain outcomes -- such as changes in attitudes -- which are not directly observable.

The major disadvantage of self-reports is their questionable *validity* -- i.e., the extent to which they provide accurate information. The validity of self-reports is the topic of considerable debate, especially when used to assess value-laden behaviors. However, self-reports constitute a useful and acceptably valid measurement methodology if a number of precautions are taken.

First, the confidentiality of responses must be closely guarded, and the procedures for protecting confidentiality emphasized to respondents. Second, the importance of the data collection and the respondents' role in the data collection should be emphasized. If the respondent feels that he or she is a key "partner" in the evaluation effort, the probability of providing valid data will be increased.

Finally, self-report data collection instruments should be carefully pilot-tested on respondents similar or identical to those who will participate in the evaluation. As part of the pilot-testing procedure, respondents should be debriefed concerning their willingness to answer all sensitive items truthfully, and suggestions should be solicited for ways that offensive items may be improved. This last may sound simplistic, but experience suggests that pilot respondents are a key source of ideas for ways to improve the validity of self-reports.

Other-Reports

Other-reports are identical in all respects to self-reports, except that data are gathered about a target individual from someone other than the target himself or herself.

One source of other-reports are individuals who have no stake in the performance of the target individual. Such other-reports may be particularly useful in assessing changes in clinical practice behavior. For example, rather than relying entirely on clinicians' reports regarding whether or not they questioned patients about alcohol and other drug use, the evaluator may conduct exit interviews with patients.

Other-reports may also be gathered from individuals who know the target individual well (spouse, other relative, close friend, etc.). Such reports may be useful in studies of patient outcomes, where patient self-reports may provide biased estimates of alcohol or other drug use

Behavioral Observations

Observations may be designed to measure some of the same behaviors assessed through self-reports or other-reports. However, observational data collection relies on the ratings or assessments of trained and unbiased observers. For example, observations of provider/patient interactions can be used to assess interviewing or counseling skills, or to assess patient reactions to the practitioner. Similarly, observation of a lecture or seminar can assess the skill of the instructor, the extent to which the sessions accomplish the objectives set forth in a curriculum guide, the nature of learner interaction, and so on. The major disadvantage of observational techniques is that the presence of the observer may change or inhibit the behavior of those being observed. Thus, the use of observers is most advantageous when it is unobtrusive.

One variant of observational data collection which overcomes some of its disadvantages is the use of simulated or standardized patients. As discussed earlier, this technique employs actors trained to present a standardized complaint. Depending on the evaluation design employed, the clinicians being assessed may either know that they are interacting with a simulated or standardized patient, or they may simply be told that such patients will be intermingled with actual patients seen by the clinicians.

Records Data

Health and mental health settings keep a variety of records on patients which may be used to assess the outcomes of faculty development and clinical training programs. Admissions records, patient charts, and insurance claims records all can be used to indicate the extent to which alcohol and other drug issues are being addressed with patients.

For example, the International Classification of Diseases, 9th Edition, Clinical Modification (ICD-9-CM) is one system used in many medical settings for the preparation of third-party reimbursement documentation or in medical record keeping. ICD-9-CM provides a standardized, detailed classification of mortality and morbidity information, as well as classification of diagnostic, therapeutic, and surgical procedures. The ICD-9 diagnostic codes relating to substance abuse cover a wide range of morbidity. Separate codes are allotted to abuse and dependence, and sub-codes within these classifications include alcohol abuse and depend-

tions of drugs including tobacco and various poly-drug combinations. Separate categories are allotted to acute intoxication, withdrawal, medical and psychiatric sequelae, fetal complications, and personal history of alcoholism. Still other codes are used to indicate substance abuse as a contributing factor to other morbidity. ICD-9 procedure codes related to substance abuse include alcoholism and drug counseling, and referral to alcoholism and drug addiction rehabilitation.

A general disadvantage of records data is that they are usually gathered for purposes other than evaluation. Thus, the data they contain may be incomplete or inaccurate, or the record keeping methods may change during the course of an evaluation. Insurance records appear to be more carefully maintained than other medical records in terms of fiscal information, but are probably subject to the same reporting errors and biases as other patient records.

QUALITATIVE EVALUATION METHODS

Data Collection Techniques

As discussed previously, certain types of evaluation questions, especially more general, exploratory questions, are most appropriately evaluated using qualitative methods. The qualitative methods most often used are observation, participant observation, and qualitative interviewing.

Observation. Behavioral observation was discussed in the previous section. Such observation can be highly structured and quantified. For example, observers might count the number of times per minute a clinician makes eye contact with a patient in order to test hypotheses concerning the quality of clinician-patient interaction. On the other hand, if the evaluator has not developed specific hypotheses, more general observations might be carried out. The evaluator might want to observe clinician-patient interactions as a way of developing a categorization of different types of intervention approaches or different types of patient reactions to interventions. Or the evaluator might want to learn something about the general tone of interactions or get an idea if and how learners are putting into practice the techniques suggested in training. Here, a structured observation protocol listing the general issues to be addressed in the observation sessions could be developed to guide the data collection.

Participant Observation. The examples of observation techniques described above relate to observers who sit quietly in the background and do not become involved. Participant observers, on the other hand become immersed in the situation, experiencing it, at least to some extent, as an actual participant would. For example, the participant observer might actually go through an educational intervention along with the other learners in order to experience first hand what the learning experience is like and to hear the reactions of other participants. Such observations would be useful in evaluating the quality of implementation of an intervention, developing hypotheses about the outcome of the intervention, etc.

Observation and participant observation can be impressionistic and even casual in initial exploratory phases of evaluation. However, these observation techniques can also be quite systematic, with the periods of evaluation precisely selected and the observations carefully documented. The usual form of documentation is detailed field notes describing the setting, the participants, conversations and events, as well as the observer's reactions. The notes should be written immediately after the event. Ample time should be allocated to the preparation of notes; it may take as much time to write the notes as it did to observe the events.

Qualitative Interviews. Qualitative interviews seek in-depth data from the respondent's perspective. The interviewer usually works from a list of general topic areas to be covered, but the questions used are open-ended. The interviewer does not structure the responses or constrain the information the respondent provides.

QUALITATIVE ANALYSIS

The purpose of qualitative research is to understand rather than to enumerate. Therefore, data collected using qualitative techniques yields a narrative analysis that can take a number of forms. For example, responses to an open ended question can be reproduced entirely, without comment. Responses can be organized by topic (e.g., all comments dealing with learners' apprehensiveness about discussing alcohol and other drug issues with patients). Qualitative data also lends itself very well to the development of descriptive case studies. When more than one case is described, they can be compared and contrasted. Frequently, the results of a qualitative study lead to more specific hypotheses and typologies that can be tested using quantitative data collection and analysis. The qualitative data can then be used to enrich quantitative reports with illustrations and examples.

CONCLUSION

The recent interest in further integrating alcohol and other drug issues into the professional training of health and mental health professionals promises to improve the health care system's response to a major cause of mortality and morbidity. Careful evaluations of faculty development and clinical training programs in alcohol and other drug abuse can help insure that the growth of these programs is planful and rational, and that the strategies shown to be successful in one institution can be exported to others.

It is our hope that the materials presented in this *Guide Book* contribute to an understanding of what evaluations can and cannot accomplish, and that interested readers will examine the sourcebooks we have recommended in their further explorations of evaluation issues.

NOTES

Adapted from French, J. and Kaufman, N. (Eds.), *Handbook of Prevention Evaluation: Prevention Evaluation Guidelines*. Rockville, MD: National Institute on Drug Abuse, 1981.

[2]Kish, L.A. *Survey Sampling*. New York: Wiley, 1967.

[3]Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press, 1969; Cohen, J. and Cohen, P. *Applied Multiple Regression Analysis for the Behavioral Sciences*. New York: John Wiley and Sons, 1975.

[4]*Ibid.*

[5]*Ibid.*

[6]Adapted from French, J. and Kaufman, N. (Eds.), *Handbook of Prevention Evaluation: Prevention Evaluation Guidelines*. Rockville, MD: National Institute on Drug Abuse, 1981.

[7]Webb, E. Unconventionality, triangulation, and inference. In N. Denzin (Ed.), *Sociological Methods: A Sourcebook*, New York: Aldine Press, 1970.